EXHIBIT 155

Page 1

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

RICHARD KADREY, ET AL.,)
Individual and Representative)
Plaintiffs,) Lead Case No.
v.) 3:23-cv-03417-VC
META PLATFORMS, INC.,)
Defendant.)

----

* * * HIGHLY CONFIDENTIAL * * *
* * * ATTORNEYS' EYES ONLY * * *
VIDEO-RECORDED 30(b)(6) DEPOSITION OF
MICHAEL CLARK (mitigation)
MONDAY, MARCH 3, 2025
DENVER, COLORADO
2:59 P.M. MST

REPORTED BY KATHY L. DAVIS, CRR, CMR

DIGITAL EVIDENCE GROUP
1730 M. street, NW, Suite 812
Washington, D.C. 20036
(202) 232-0646

Page 43

repetitive and duplicative." Repetitive was the first bullet.

Q Mm-hmm.

A And then duplicative and repetitive are kind of the same.

Q Mm-hmm.

A My phrase in my prior testimony feels incomplete, but please ask your question.

Q So for the examples of filter data and repetition, he provides copyright information as repetitive data, he provides personal identifying information as repetitive data, right?

A We did miss a few bullets, Remove excessive new line crack terms and Remove documents in the token distribution tool.

Q Mm-hmm.

A Um -- I'm still unclear on the question. I apologize.

Q My question is that when you're talking about this script that he wrote to remove data that had a higher likelihood of being regurgitated, among the repetitive data that he removes is also -- it's -- is copyright data; is that correct?

A Um, among the data that's being removed, the repetitive and duplicative data has a higher

Page 44

likelihood of being regurgitated. The copyright information and line spacing and emails, part of that is for preventing regurgitation, but also just general stability of the model from -- from a language perspective.

Q Okay.

MS. POUEYMIROU: Can we turn to Tab 15. This will be 14, Exhibit 14.

(Clark Exhibit 14, marked for identification.)

Q (BY MS. POUEYMIROU) Are you familiar with this document, Mr. Clark?

A I don't remember this document.

Q Are you familiar with Meta's source metadata project?

A Let me read this to see if I am.

MS. POUEYMIROU: How much time?

VIDEOGRAPHER: 50 minutes.

MS. POUEYMIROU: 5-0?

VIDEOGRAPHER: (Nodded.)

MS. POUEYMIROU: Thought it was 10. Almost gave me a heart attack.

(A discussion was had off the record.)

A Okay.

Q (BY MS. POUEYMIROU) Okay. So the

Page 45

concern about source metadata leaking leads to Meta employees talking about whether it can be hashed or unhashed in use cases; is that correct?

A More than just source metadata leaking as potential harm. What's -- what's the question?

Q My question is, they're trying to figure out a safe way to train and source metadata without risking IP or brand safety. Is that -- is that fair to say?

MR. WEINSTEIN: Object to form.

A With the -- going back to Exhibit 13, it is much more than just leakage of info. It is also training stability and performance, that what does either the hash or the raw metadata introduce from a -- from a instability perspective into the model?

Q (BY MS. POUEYMIROU) Okay. I recognize that, but another consideration is about IP and brand safety; is that correct?

A That is one of the three that is listed.

Q Okay. And so the document that you now have before you -- this is Exhibit 14 -- is entitled Source Metadata Strategy. Do you see that?

A I do see that.

Q And I had some questions about that. First, are you familiar with this source metadata

Page 46

project at Meta?

A Personally, I was familiar that this was a -- this was work that was being evaluated.

Q And why were you personally familiar with that?

A Just in -- working in generative AI and seeing other -- other projects that were in flight is how I'm familiar -- or were in the works is why I was familiar with it.

Q And how did you understand the project?

A I understood the project as, Could adding source metadata to the model help validate where we see issues or regression in certain types of performance, know where that came from, but also from an academic perspective in trying to get to a place to where as -- the question of, Was Abraham Lincoln a president, where that knowledge came from.

Q So the -- the metadata that Nikolay Bashlykov had stripped from LibGen, for example, was metadata that Meta is now considering how it can safely train on; is that correct?

MR. WEINSTEIN: Object to form.

A The data that Mr. Bashlykov's script removed was a subset of the overall data that was being tested and evaluated as part of the source

Page 47

metadata strategy. However, in looking and comparing this data, the data in Mr. Bashlykov's were where lines contained that --

(Weather noise was heard.)

THE DEPONENT: Is that all wind?

MS. POUEYMIROU: It sounds like it.

MR. WEINSTEIN: Okay.

A Okay. Apologies. To get back to answering the question, what Mr. Bashlykov stripped out were individual lines --

Q (BY MS. POUEYMIROU) Mm-hmm.

A -- but not actually the full set of metadata or source metadata that would have been needed to have filled in -- and if I look at the example, Title, Type, Host Name, URL, Date, Generator, Source, Media --

Q Mm-hmm.

A -- Media Bias, and other information to fill out the metadata fields.

Q So what's the URL when you are -- you've gotten your data from LibGen? Will that be in the URL?

A This was -- the section I just read was specific to Common Crawl.

Q Okay. And so if you go to the next page,

Page 53

documents from other URLs?" What is this concern about people figuring out the hash associated with Wikipedia?

MR. WEINSTEIN: Object to form.

A I would have to speculate that the harm is going back to releasing information about the data mix that was used to build out the model. I believe that comment was from Eric Smith, not Soumya.

Q (BY MS. POUEYMIROU) Oh, it looked like it was on the back.

A Oh, I was on 81207.00005.

Q Yeah.

A Okay. Eric Smith had made that comment, just for accuracy.

Q Oh, okay.

So the leakage of source metadata exposing to the public what Meta had trained on --

MR. WEINSTEIN: Object to form.

A The proprietary data mix that was used for developing the model.

Q (BY MS. POUEYMIROU) And what do you mean by "proprietary data mix"?

A The combination of data that was developed -- the combination of data -- the exact combination of data that was used to train the model.

Page 54

Q How would -- how would a person know the combination or the data mix of your data --

MR. WEINSTEIN: Object to form.

A I --

Q (BY MS. POUEYMIROU) -- as opposed to just data that was trained on?

MR. WEINSTEIN: Object to form.

Q (BY MS. POUEYMIROU) I have always thought of data mix in this case as how much of one data mix compared -- one dataset compared to another. Are you just talking about the data that Meta trained on being released to the public?

A The com- -- the exact combination of what datasets and/or data Meta trained on for that model.

Q How does the -- how does the training on the title and author of a work --

MR. WEINSTEIN: Object to form.

MS. POUEYMIROU: I didn't finish.

MR. WEINSTEIN: Oh, sorry. Sorry about that.

Q (BY MS. POUEYMIROU) -- expose a data mix?

MR. WEINSTEIN: Object to form, scope.

A The -- only the -- as the motivation said, the reason that you would train on it is to

Page 55

make it available from an explainability or to improve the quality from a factuality perspective.

And so there's no sense in training -- in adding additional data that can't be used, which therefore exposes that data.

Q (BY MS. POUEYMIROU) So I'm trying to understand, though, your point about proprietary datasets. Are you saying that LibGen is Meta's intellectual property and if training on source metadata from LibGen, which then gets regurgitated, it would expose?

MR. WEINSTEIN: Object to form.

A Just to -- I believe that's a misclarification of the phrase that I used. It's the proprietary data mix, the combination of what data was used to train not the data itself being proprietary.

Q (BY MS. POUEYMIROU) So you're saying the fact that Meta trained on Common Crawl and Wikipedia and Substack and Archive and . . . You're saying that is the concern, that that would be released to the public?

MR. WEINSTEIN: Object to form.

A That would be a -- that is a potential example, yes.

Page 56

Q (BY MS. POUEYMIROU) But didn't Meta speak openly about its use of Common Crawl and other datasets in its published papers?

MR. WEINSTEIN: Object to form. Outside the scope.

A It only spoke about that in LLaMA 1, which was a research release only, and not in any other models after. That was the one mitigation that I forgot in my list when I read them off, was that -- to not share that data mix, to not share the data mix as part of one of the mitigations.

Q (BY MS. POUEYMIROU) Okay. And so your testimony is that by removing the source metadata, Meta was able to conceal what it trained on because by training on source metadata, the regurgitation risk was higher?

MR. WEINSTEIN: Object to form.

A I believe that's a mischaracterization of my testimony as the source metadata document that was listed here was adding additional data into the training mix and not taking it out.

Q (BY MS. POUEYMIROU) Was that --

A And --

Q -- because it had first been taken out?

MR. WEINSTEIN: Object to form.

Page 57

A This -- sorry. I wasn't complete with my answer. Do you mind asking the original question?

Q (BY MS. POUEYMIROU) Sure. The decision to add source metadata to training is predicated on the fact that Meta had already removed source metadata --

MR. WEINSTEIN: Object to form.

Q (BY MS. POUEYMIROU) -- from training; is that right?

MR. WEINSTEIN: Sorry. Object to form. Lacks foundation.

A If I look at the examples from -- which were in 65250 from what Mr. Bashlykov had removed, this information is incomplete in how it's structured and was not trained in a structured way that could have been used in the same way as the source metadata strategy. This was removing specific things where they were placed in the doc that were not structured or in a way that could actually help what the motivation of the test that was being done around the source metadata strategy was, which, to the best of my knowledge, was not actually pursued because of challenges it created in both the performance and stability of the model when it was added.

Q (BY MS. POUEYMIROU) And also with

Page 58

respect to --

A The reasons it was not chosen is it created too much regression, added too much additional volume, which you can see reflected in the comments of data that wasn't actually usable as part of inference and was therefore not pursued because of the challenges around performance and stability of the model.

Q You're saying that with respect to personal identifying information too? That was why it was removed?

MR. WEINSTEIN: Object to form.

A We were just talking about the source metadata strategy project, which was explicitly adding multiple structured fields into the model. The PII that was removed as part of Mr. Bashlykov's script was to prevent repetitive personal information from being able to be regurgitated about private individuals. That might have been email or other kinds of PII.

Q (BY MS. POUEYMIROU) And the use of that script to remove the copyright information was also a regurgitation concern?

MR. WEINSTEIN: Object to form.

A That was both regurgitation, plus the

Page 59

repetitive characters of copyright and the C -- the copyright mark -- copymark would affect the performance of the model as well because that would be very much repetitive and become part of outputted responses, but not necessarily adjacent to copyrighted information.

Q (BY MS. POUEYMIROU) Okay.

MS. POUEYMIROU: Can we pull out 23.

(Clark Exhibit 15, marked for identification.)

A Are we switching topics?

Q (BY MS. POUEYMIROU) We're switching documents.

A Okay.

Q The topic is mitigation, though. Now, this is a multipage document. I'm only going to be asking about parts later on, so just take your time.

MS. POUEYMIROU: This is Exhibit --

THE REPORTER: 15.

Q (BY MS. POUEYMIROU) -- 15. All right. I want to look at the page ending in 2225. Do you want to take a look at that page?

MS. POUEYMIROU: How much time do we have?

VIDEOGRAPHER: 24 minutes.

Page 75

STATE OF COLORADO)
)ss. REPORTER'S CERTIFICATE
COUNTY OF DENVER)

I, Kathy L. Davis, do hereby certify that I am a Registered Professional Reporter within the State of Colorado; that previous to the commencement of the examination, the deponent was duly sworn to testify to the truth.

I further certify that this deposition was taken in shorthand by me at the time and place herein set forth, that it was thereafter reduced to typewritten form, and that the foregoing constitutes a true and correct transcript.

I further certify that I am not related to, employed by, nor of counsel for any of the parties or attorneys herein, nor otherwise interested in the result of the within action.

In witness whereof, I have affixed my signature this 4th day of March, 2025.

Kathy L. Davis

Kathy L. Davis
Certified Realtime Reporter